DOCUMENT RESUME

ED 166 684 CS 004 696

AUTHOR Wolf, Anne E.; Greenwald, M. Jane
TITLE Developing an Observation Schedule for Coding Reading-Related Activities in the Secondary School Classroom.
PUB DATE 78
NOTE 9p.; Paper presented at the Annual Meeting of the National Reading Conference (28th, St. Petersburg Beach, Florida, November 30-December 2, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Classification; *Classroom Observation Techniques; *Classroom Research; *Content Reading; *Evaluation Methods; Language Arts; Reading Instruction; *Reading Research; Secondary Education; *Teacher Behavior

ABSTRACT

A study was undertaken as part of the process of trying to develop an instrument for systematically coding reading and other language arts activities in the secondary content area classroom. Six teachers chosen from five secondary schools were each observed while teaching four randomly selected classes over a two-week period. All observations were audiotaped and nonverbal activity was recorded in a log for use later by coders. Of the 22 categories in the schedule, 13 proved to be too narrow or too uncommon to discriminate among the teachers. In three other categories, behaviors occurred frequently but were so similar as to fail to distinguish among teachers. Six behaviors occurred with enough frequency and variety among teachers to justify calculating generalizability (reliability) coefficients for the categories. Data from this study suggest that reading and listening activities do occur in the content classroom with enough frequency and variety that reading and listening activity categories can be devised that will be generalizable. (TJ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Anne E. Wolf
Assistant Professor
Department of Secondary and Continuing Education
Brewer Hall Room 154
University of Wisconsin
Eau Claire, WI 54701

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Anne E. Wolf
M. Jane Greenwald

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED166684

DEVELOPING AN OBSERVATION SCHEDULE FOR CODING READING-RELATED ACTIVITES IN THE SECONDARY SCHOOL CLASSROOM

Anne E. Wolf
University of Wisconsin-
Eau Claire

M. Jane Greenewald
University of Wisconsin-
LaCrosse

A paper presented at the annual meeting of the National Reading Conference, St. Petersberg, Florida, December, 1978.

Many secondary methods texts recommend that content-area reading assignments be linked to other language arts activities such as oral or written previews of difficult vocabulary. Review of the literature indicates that virtually no researchers have systematically observed instruction in secondary classrooms to determine the variety, extent of use, or effectiveness of reading and reading-related instruction. Dolan (1974) investigated the extent and context of reading activities in British primary and secondary classrooms, and observation in American classrooms has concentrated on reading instruction in grades K-6 (Coker & Lorentz, 1977; Erlich & Borich, 1976; Hautala & Aaron, 1977; Jerrolds, et al., 1976; Mallula, 1978; Quirk et al., 1973; Quirk, et al., 1976). While

CS004696

these studies are of methodological interest, instruments and results are not applicable to secondary instruction in American schools.

Wolf and Greenewald (1978) developed an observation schedule for systematically coding reading and other language arts activities in the secondary content-area classroom. The instrument consists of 22 categories grouped into 5 language arts clusters: students read, students listen, students write, students discuss, and other. The 10 reading categories include straight reading, reading co-occurring with a listening activity such as text explication, and reading co-occurring with a performance activity such as a lab exercise. This study reports on the reliability of the instrument's scores.

Reliability was estimated within the framework of Generalizability Theory (Cronbach, et al., 1963; Cronbach, et al., 1972). Classroom observation schedules are subject to more potential sources of unreliability than paper-and-pencil tests. These include infrequency or instability of teacher behavior as well as rater inconsistency. While most observational studies report interrater consistency alone (McGaw, et al., 1972), infreqency and/or instability of teacher behavior has also been an important source of instrument unreliability (Calkins, et al., 1977; Erlich & Borich, 1976). Generalizability theory is a more appropriate framework for conceptualizing and estimating the reliability of an observation schedule than is traditional reliability theory because it allows the researcher to explicitly and simultaneously estimate more than one source of score variation (Borich, et al., 1977; Calkins, et al., 1977; Erlich & Borich, 1976). Research on classroom observation schedule reliability indicates that generalizability can be difficult to achieve (Borich, 1977a; Borich, 1977b; Calkins, et al., 1977; Shavelson & Atwood, 1977).

Method

Data were collected in the spring of 1978. Teachers in 5 secondary schools in western Wisconsin were asked to participate in the study, and 6 volunteers were selected. Each teacher was observed teaching 4 randomly selected classes over a

two-week period. Teachers were not informed beforehand of which classes would be visited or the days of the visits. All observations were audiotaped, and nonverbal activity was recorded in a log for use later by the coders. Three graduate students in education at the University of Wisconsin-Eau Claire independently coded the data after participating in a 3-hour training session.

Variance components and generalizability (reliability) coefficients were calculated first for each of the observation schedule's 22 categories and then for the five language-arts clusters. Variance components estimate the spread of scores in the population for each source in a design and are used to calculate the generalizability (reliability) coefficients that estimate score stability. Univariate analysis of variance for a three-factor partially nested design with a total of 72 degrees of freedom allowed for estimation of variance components attributable to teachers, to raters, to occasions nested within teachers, to teacher by rater interaction, and to residual.

Results

Scores for different teachers are generalizable when real differences among them are greater than differences due to unwanted variation from sources such as occasions or raters. Low generalizability therefore may result from either too much unwanted variation or from too little real variation--either because teachers in fact do not vary or because the behavior is rarely observed. Of the 22 categories in the schedule, 16 lacked generalizability because of very small or nonexistent teacher variance components indicating little variation among teachers in the population. Inspection of the raw data indicated that in 13 of these 16 categories the target behavior was coded less than 25% of the time. These categories were either too narrow or too uncommon to distinguish among different teachers. In 3 of the categories, on the other hand, the behaviors did occur frequently: listening to class procedures (77% occurrence); listening to reviews

and previews (46% occurrence); and listening to references to reading (51% occurrence). Behavior within these categories was frequent but so similar that scores also failed to distinguish among teachers. Teacher means for the listening to class procedures category, for example, ranged from 2.3% of class time to 6.9% of classtime.

In this study the generalizability coefficient was a ratio of a teacher variance component to the sum of the variance components for teacher, occasions nested within teachers, teacher by rater interaction, and residual. Six behaviors occurred with enough frequency and variety among teachers to justify calculating generalizability (reliability) coefficients for the categories. These generalizability coefficients were .17 (reading co-occurring with listening); .30 (reading co-occurring with writing); .21 (reading co-occurring with two or more language arts activities excluding the combinations reading/reading/writing and reading/writing/discussion which appeared in other categories); .38 (reading co-occurring with performance such as reading a lab manual to carry out a chemistry experiment); .40 (listening to content); and .18 (activities unrelated to content such as attendance taking or listening to school announcements). These coefficients all fell below an acceptable .70 level. In order for these six categories to be generalizable for this population, more than six teachers would need to be observed on more than four occasions.

Proportions of total variance were calculated for the variance components in each of the 6 categories to estimate the relative effect of each source on total expected score variation. Across the 6 categories, proportions ranged from 17 to 39% for teachers, from 9 to 5.8% for raters, from 3.7 to 61% for occasions nested within teachers, from 0 (based on a negative component) to 16% for teacher by rater interaction, and from 14 to 40% for residual. While the proportions of total expected variance varied from category to category, these data suggest that variation due to teachers, occasions, and residual generally accounted for much

more of the total population score variation than did variation due to raters and to the rater by teacher interaction.

Because there were few or no observations for many of the schedule's categories, the original 22 categories were collapsed into the five language arts clusters and the data reanalyzed. The new categories were not more productive than the originals.

Conclusions

The preliminary observation schedule evaluated in this generalizability study is in need of substantial revision: most categories encode behaviors that occurred too infrequently to be generalizable; a few encode behaviors that occurred frequently but too consistently to discriminate among teachers; and 6 exhibited sufficient generalizability to warrant refinement and further testing in a revised schedule. Data from this study suggest that reading and listening activities do occur in the content classroom with enough frequency and variety that reading and listening activity categories can be devised that will be generalizable.

Rater effects generally represented a small proportion of total expected score variation, suggesting that interrater inconsistency is a relatively minor contributor to lack of generalizability. Raters in this study received minimal training. Interrater reliability in future studies should be as high or higher if raters are given as much or more extensive training. In any case, the small percentages of total variance due to raters suggest that further classroom observation studies can be conducted with a minimal number of raters. These results are consistent with the findings of other generalizability studies on observation schedules that interrater reliability is a relatively unimportant source of inconsistency in observation schedule scores.

Occasions within teachers contributed substantially to score variation in the population. Teachers do vary across occasions, though considerably more so in some categories than others. Again, this conclusion is consistent with those of other researchers and is to be expected in studies of teacher behavior. It is likely that more than four occasions will be needed to reliably observe at least some reading-related activities of teachers.

The relatively large residuals include random error, rater by occasion interaction, teacher by rater by occasion interaction, as well as the effects of any systematic variation not partitioned out by the design of the study. It is likely that this residual was enlarged by the fact that teachers were drawn from different content areas and grade levels and were observed teaching different classes.

As previous research has suggested, the difficulties encountered in this study are not uncommon. While it may be difficult to develop an observation schedule consisting of behaviors that are generalizable when a reasonable number of teachers are observed on a reasonable number of occasions, the question of what is occurring in secondary classrooms appears important enough to warrant further work on instrument development. Results of this study suggest that a reliable schedule can be developed.

REFERENCES

Borich, G. D. The appraisal of teaching: concepts and process. Reading, Mass.: Addison-Wesley, 1977a.

Borich, G. D. Sources of invalidity in measuring classroom behavior. Instructional Science, 1977b, 6, 283-318.

Borich, G. D., Malitz, D., Kugle, C. L., & Pascone, M. Convergent and discriminant validity of five classroom observation systems: testing the model (Report 5045). Dallas: The Research and Development Center for Teacher Education, 1977.

Calkins, D. S., Borich, G. D., Pascone, M., & Kugle, C. L. Generalizability of teacher behaviors across classroom observation systems (Report 5046). Dallas: The Research and Development Center for Teacher Education, 1977.

Coker, H. & Lorentz, J. L. Observed patterns of teacher-pupil classroom behavior as predictors of student growth in reading. Paper presented at the meeting of the National Reading Conference, New Orleans, December, 1977.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: a liberalization of reliability theory. The British Journal of Statistical Psychology, 1963, 16 (Pt. 2), 137-163.

Dolan, T. The incidence and context of reading in primary and secondary schools. In D. Moyle (Ed.), Reading: what of the future? (Proceedings of the eleventh conference of the United Kingdom Reading Association, Edge Hill College, 1974.

Erlich O. & Borich, G. D. Generalizability of teacher process behaviors during reading instruction (Report 5042). Dallas: The Research and Development Center for Teacher Education, 1976.

Hautala, L. W. & Aaron, R. L. Time-Use: a variable in teacher effectiveness. Paper presented at the meeting of the National Reading Conference, New Orleans, December, 1977.

Jerrolds, B. W., Beton, S. E., & Bazemore, J. An instrument designed to identify specific strengths and weaknesses in reading instruction. In W. D. Miller & G. H. McNinch (Eds.), Reflections and investigations on reading, Twenty-fifth Yearbook of the National Reading Conference. Clemson, S. C. National Reading Conference, 1976.

Mallula, S. A. Teacher competency, teacher attitude, and student attitude toward reading. Unpublished doctoral dissertation, University of Georgia, 1978.

McGaw, B., Wardrop, J. L., & Bunda, M. A. Classroom observation schemes: where are the errors? American Educational Research Journal, 1972, 9, 13-27.

Quirk, T. J., Trismen, D. A., Weinberg, S. F., & Nalin, K. B. The classroom behavior of teachers and students during compensatory reading instruction. Princeton, N.J.: Educational Testing Service, 1973.

Quirk, T. H., Trismen, D. A., Weinberg, S. F., & Nalin, K. B. Attending behavior during reading instruction. The Reading Teacher, 1976, 29, 640-646.

Shavelson, R. & Atwood, N. Generalizability of measures of teaching process. In G. D. Borich, The appraisal of teaching. Reading, Mass: Addison-Wesley, 1977.

Wolf, A. E. & Greenewald, M. J. Observation schedule for coding reading-related activities in the content classroom. Unpublished manuscript, 1978. (Available from Dr. Anne Wolf, University of Wisconsin-Eau Claire).